Plan Overview

A Data Management Plan created using DMPonline

Title: HUGODECA

Creator: Laurence Noel

Principal Investigator: Laurence NOEL

Data Manager: Laurence NOEL

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0001-7577-3794

Project abstract:

The single ground-breaking goal of the HUGODECA project is to describe the cellular composition and organization of the developing human gonads and to understand how it changes during sex determination into testes in males and ovaries in females. What are the underlying mechanisms and first molecular and cellular events that accompany the differentiation and divergence of embryonic gonads? How and when do male and female cell lineages diverge and specific traits emerge? HUGODECA focuses on healthy gonad development, but our reference model will be tested using specific ex vixo assays mimicking Differences/Disorders of Sex Development (DSD). To reach this ambitious goal, the HUGODECA consortium brings together leading European academic and industrial experts (from 5 different EU countries). It includes active contributors and executives of the Human Cell Atlas (HCA) which will ensure complementarity with other ongoing HCA efforts. The overall HUGODECA concept is grounded on the integration of multiple synergistic expertise and technologies: Single cell profiling, Spatial transcriptomics, 2D Mass cytometry and cyclic immunofluorescence and 3D imaging of optically cleared gonads. HUGODECA will implement novel tools, analytical and computational methods to process and integrate multidimensional OMICS and image data across different platforms. It will evaluate the accuracy of ex vivo culture models of human gonadal development and assess consequences of altering key signaling pathways. HUGODECA will build the first multiscale developmental cell atlas and unprecedented reference maps of male and female human gonads. It will implement an interactive and multi-dimensional online portal for clinicians, scientists and public including DSD patient associations. HUGODECA shall improve our understanding of DSD, which is a major pediatric concern, requiring complex and highly specialized medical treatment and psychosocial care.

ID: 55010

Start date: 01-01-2020

End date: 30-06-2022

Last modified: 02-05-2022

Grant number / URL: 874741

Copyright information:

HUGODECA - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

The goal of the *HUGODECA* project is to describe the organization of the developing human gonads and to understand the molecular and cellular mechanisms underlying sex determination. In order to do so, the project aims to make major advances in the following complementary areas:

- scRNA- and ATAC-seq of human fetal gonads, with assessment of cell surface markers by flow cytometry (WP1);
- ST and ISS leading to the generation of 2D and 3D expression maps of targeted genes at single-cell resolution (WP2);
- Spatial proteomics based on diverse imaging methods (WP3);
- Software development to enable the integration of large OMICS data with imaging data (WP4);
- Organotypic culture models to evaluate the consequences of signaling pathways directing sex-specific differentiation and endocrine disruptors on gonadal development (WP5).

This project will lead to the generation and process of multiple types of research data, that is data produced as the research is conducted, and which should be as re-usable as possible by the scientific community. Most of their characteristics, such as format and size, vary according to the techniques used to generate them. For ex vivo experiments, the input and the output of the assay are both biomaterials; for imaging techniques, the input will still be some biomaterials, but the output will be an image. In the case of a single-cell experiment, the data first generated correspond to a textual sequence referring to the genetic code. The frontier between the different types of data generated or processed, however, is not always so clear-cut and one of the stated goal of the HUGODECA project is actually to be able to integrate heterogeneous types of data into one common scaffold. For this reason, it will sometimes be useful to be able to distinguish the primary medium of information from the complementary media, which have been generated to describe the primary one. In the case of ST, for instance, the image is the primary medium, and OMICS expression data will be used as a means of giving complementary information on that medium. Actually, the way these complementary data is stored may also vary: they can be mentioned in a separate file that will be part of a data package, or correspond to embedded metadata, integrated in the primary media.

Some of the complementary data which can be found in sets of research data are specially intended to document the way experiments are carried out:

- Study, sample and assay metadata: information about the study, the materials used (from biomaterial to lab equipment) and the conditions of the experiment;
- Methodological data: the protocol in itself, with the different steps of the procedure, and/or descriptions of data workflows;

As diverse as the technologies involved may be, these research data also share some common process steps and we have classified them according to the following sub-types:

- Raw data, resulting directly from the technology with which they have been acquired (initial image captures or sequencing data, for instance);
- Pre-analysis data: raw data having undergone a transformation process deemed necessary to be able to proceed to a meaningful analysis or generated in order to be able to run the main analysis. As we will mention later on, these types of data are generally not kept in the long-term;
- Analysis data: data associated to quantitative and/or qualitative features;
- Interpretation data: lists of biological elements, which have been identified as being of interest, or computational data resulting from the interpretation of the analysis.

When considering data from the point of view of the process, another specific type of data comes into the picture: coding data, from scripts written to establish a process pipeline to the complete source code of a software solution specially developed for the project.

Finally, the HUGODECA project will also lead to the production of structured documents, such as publications and communication supports, with the obvious goal of sharing and promoting the results achieved by the HUGODECA Consortium among the scientific community.

In the following subsection, we have listed the different types of expected data by task, to get a more detailed view of their specific characteristics (as the production of publications and communication supports is not task-dependent, they are not mentioned in this list but we will discuss the question of their findability and accessibility in the next sections). All the data mentioned below are not intended to be published or kept in the long-term: the intent, here, is more to have a broad view of the different data generated or processed in the short-term, so as to be able to take them into account when defining naming conventions or characteristics such as the data volume. It is to be noted that this last point is a sensitive issue for the project as the technologies involved imply to store several TB of data (even for a single task) which make the processes of data storage, transfer, back-up and preservation more difficult to address than for datasets of smaller sizes.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Study and assay metadata will be created to document the different experiments defined in WP 1, WP 2, WP 3 and WP 5 (this is not applicable to WP 4, since it deals with software development).During the first stage of the project, we will gather information about the program and the different studies that are planned by using SEEK (D7.1). This application follows the Investigation, Study and Assay (ISA) specifications, and we will create one investigation by work package, and one study by project task.To become findable (and citable), datasets should get a unique, persistent identifier. Persistent identifiers like Digital Object Identifiers (DOI) have been designed as a solution to avoid link rot (that is, when a hyperlink stops referring to the original source because it was moved). These persistent identifiers thus ensure that the data is and will be findable. The use of SEEK will make it possible to generate a DOI for each study.
When it comes to versioning :

- **For documents** (.doc, .pdf, ...), versioning will be maintained by following this pattern :[document name]_[version number]_[status: DRAFT{0,1}]. Example : HUGODECA_DMP_7-4_V1.doc
- **For data that can be generated multiple times**, the creation date may be indicated instead of a version number. The date will then be expressed with this format: YYYY-MM-DD
- **For software and scripts**, versioning will be controlled via the use of dedicated tools (Subversion, Git...)

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Datasets generated and processed by the HUGODECA consortium intend to be listed on the HUGODECA web portal (D6.4), and the links to download the datasets will then point towards the different open repositories. As already mentioned, the consortium members intend to take part in the HCA initiative, so analysis data may also become available on the HCA data portal.
For Software and tools, Genomics and transcriptomics data are in file formats which are widely used within the scientific community, that is fastq.gz files for raw sequencing data, bam files for aligned sequences and count matrices in HDMF5-loom, mtx or csv formats for processed data, making these fully readable and re-usable worldwide from a technical viewpoint. Raw images can be viewed with open-source softwares like the Fiji/ImageJ tools (Fiji is an open source project hosted in aGit version control repository and comprehensive documentation, ImageJ source code is available online, with its user manual). Open-source softwares likeCell Profiler and histoCAT can also be used to analyse imaging data. However, in order to visualize the fully annotated 2D and 3D cell maps, Keen Eye 3D viewer will be necessary: Keen Eye Platform™ is a proprietary and patented software-as-a-service platform and the viewer will be accessible through a cloud service, promoting high accessibility across a worldwide research community.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

Interoperability denotes the ability of diverse systems and organizations to work together (inter-operate), to “plug together” components in larger complex systems. It is a key aspect to consider if the participants involved want to avoid the “Tower of Babel” effect, that is to avoid the breakdown of the tower-building effort because the terms used to communicate are not shared and understood by all. For that reason, this section will define what values those metadata can take to ensure the use of a common vocabulary among researchers, but also what types of metadata are required to make sure that the datasets may be accepted in other repositories.

Types of metadata required

For samples and processes, a special attention will be given to the metadata used within the HCA initiative.

Regarding samples, there should be metadata about:

- The organism donor (embryo) : unique anonymous identifier (as transmitted by the provider), biological sex (female, male, mixed or unknown), developmental stage (in PCW), species (“Homo sapiens”, NCBItaxon:9606), known diseases (it should default to "no disease", which is equivalent to "MONDO_0000001");
- The sample in itself: unique identifier and organ from which the sample was collected.

Regarding sample processing and data acquisition, the metadata depend on the technology used. When single-cell technologies are involved, metadata should include information about:

- The dissociation procedure
- The library preparation protocol :
  - Library construction method (e.g. “10x v3 sequencing”), library construction kit (name, manufacturer and catalogue number)
  - The input nucleic acid molecule (e.g. “polyA RNA")
  - The nucleic acid source (e.g. “single cell”)
  - End bias (e.g “3 prime tag", "3 prime end bias", "full length”)
  - Strand ("first", "second", "unstranded" or "not provided")
  - Optional properties from those listed in themetadata for HCA
- The sequencing protocol :
  - Instrument used for sequencing and model (e.g. Illumina HiSeq 4000), the sequencing design (e.g. 2x100 bp) and the number of raw reads
  - 10x specific metadata such as fastq creation method (e.g. “Cellranger mkfastq” or “Illumina bcl2fastq”), fastq creation method version (e.g. “Cellranger 3.1”)

When imaging technologies are involved, metadata should include information about:

- Tissue preparation and the imaging preparation protocol
  - Fixation
  - Image slide thickness
  - Final slicing method (cryosectioning) …
- The imaging protocol :
  - Microscopy (« fluorescence microscopy »)
  - Magnification, Numerical aperture, pixel size …

These metadata can be found in the HCA metadata dictionary. Those which are the most relevant for the project are Specimen from organism, Imaged Specimen, Imaging preparation protocol, Library Preparation Protocol, Sequencing protocol.

For biological imaging data, the OME Model defines a set of metadata to include, such as XYZ dimensions and pixels type, as well as extensive metadata on, for example, image acquisition, annotation, and regions of interest (ROIs).

Metadata values

In order for all participants to use the same terms, ontologies and controlled vocabularies are going to be used. Whenever possible these values should correspond to the terms defined in the HCA ontology. If this ontology is incomplete for our needs, the concepts used will preferably come from other existing ontologies, including but not restricted to:

- EFO (Experimental Factor Ontology) which provides information about many experimental variables available in EBI databases, while Eagle-i resource (ERO) is an ontology of research resources such as instruments, protocols, reagents, animal models and biospecimens
- the Biological Imaging Methods Ontology (fBBi) is an ontology dedicated to the terms used in biomedical research for imaging and visualization methods, and the BioAssay Ontology (BAO) describes biological screening assays and their results including high-throughput screening (HTS).
- Embryonic structure is described in UBERON and EHDAA2 provides a structured controlled vocabulary of stage-specific anatomical structures of the developing human and is linked to HSAPDB, which includes carnegie stages.
- Different ontologies describe properties and classes at the cell level : the cell ontology (CL) is a general ontology which applies to cell types in animals; and CPO structures the vocabulary related to morphological and physiological phenotypic

characteristics of cells, cell components and cellular processes.

For the different metadata fields which are required, we will thus define the ontology or controlled list from which the values should be part

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

By giving access not only to raw and analysis data, but also to study metadata, methodological data, and scripts used to facilitate the process and analysis of the data, the HUGODECA consortium pave the way for data re-usability. The fact that the different protocols established by the consortium members will be published will enable to check that quality assurance processes are ensured. In the case of coding data, documentation and literary programming is promoted : whenever possible, tutorials in the form of notebooks (jupyter notebooks, R notebook, KnitR, Sweave) will be published to document analysis scripts (potentially with the use of binder), so as to integrate the code with the corresponding narrative and documentation. The use of container technology and virtualization, such as Docker, is also encouraged.

A CC BY license (requiring only attribution) will be the preferred option for scientific publications and reports: data users will be free to “reuse the material in any medium or format, and to remix, transform, and build upon the material, even commercially” but they will be reminded that they should cite the dataset and acknowledge the data producers in any publications and presentations that make use of the data.

In conformance with the principle of being "as open as possible, as closed as necessary”, the HUGODECA consortium will also support partners in protection of their results, and help them secure future exploitation opportunities. Project members will implement a continuous and integrated strategy for monitoring *HUGODECA* results and the Innovation Management Committee will define those which are likely to be exploited, thus kept confidential. During the preliminary exploitation plan of the HUGODECA project, the deliverables D1.1, D 2.3, D4.1, D4.2, D4.3 have already been identified as potentially patentable.

In the case of bioimaging data (3.3) all requests for access will be vetted through an MTA, under the hospices of Fondation Voir et Entendre (Vision Institute, SU, INSERM) and signed by the legal officer of the receiving institution hosting the requesting PI user. Commercial destinations will be prohibited.

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

12 person-months have been requested by P01c-INSERM for data management (WP7). A data manager (L. Noël, P01c-Inserm) has been recruited to this end. She will help manage the data with the help of a data contact person, identified for each task (7 person-months requested globally to this end). They will notably implement tools necessary for data management, provide the relevant metadata during the course of the project, keep track of the dissemination level of the different data sets, and refine the naming convention.

20 000 euros have also been requested to cover the costs related to open access publications (author-publishing fees).

Cloud data storage (100 TB / month over 18 months) for 3D images has been estimated at a cost of 54000 euros: this sum has to be taken into account when considering the long-term preservation of imaging data, since this contract will have to be renewed. For that reason, partnerships with European institutes (CINES; CC-IN2P3, EU Image Data) that can provide long-term support for data hosting are under negotiations, all as well as the possibility to deposit the images in IDR.

For OMICS data, the deposition of raw and analysis data in public repositories remain the best alternative to ensure their long-term preservation: those are trustworthy data repositories, with a certificate or explicitly adhering to archival standards.

The long-term preservation of HUGODECA data is particularly valuable since they deal with unique specimens, analyzed at crucial developmental stages. As the ethical aspects are particularly sensitive, these datasets will be a major source of information, if future regulations should come to restrict this work as in other major countries. The HUGODECA programme is already of a high value for researchers who can not carry out experiments on this type of specimens due to restrictive regulations in their home country.

The data generated for the HUGODECA project have also a high commercial and educational potential: the study of the cell atlas

may eventually be included into university curriculum (medical, biology, etc.) and converted as a unique learning 3D tool for anatomy/surgery. It is also to be noted that the consortium members have already received many requests from documentary makers (TV, etc.) for anatomy content and human fetus discovery.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Initial storage of the data will be performed locally (with the exception of HUGODECA_DS_3-3, HUGODECA_DS_5-1, HUGODECA_DS_5-2, HUGODECA_DS_5-3, HUGODECA_DS_5-4 imaging data, for which a cloud data storage solution has been chosen). For academic groups involved in the project, institutional ICT facilities will ensure a secure environment, with firewall system in place, virus intruder protection, and access to digital files controlled with encryption and/or password protection and SSL encryption for data transfer. Industrial partners will observe the same types of security measures, and dedicate specific storage space for the data generated for this project.

For file transfer and backups, checksums will provide a simple way to compute the integrity of data files before and after file operations.

The sheer size of expected data make it difficult to follow back up best practices (at list 3 copies on at least two different media). As a general rule, we remind here that raw data are usually considered to be the master copy of any given record (or golden copy): therefore, there should be a back-up copy of all raw data. The use of the Dell EMC Isilon storage system (approx. 540 TB) (funded by INSERM) to store their raw data on at least one second media is under study. For images, the important point is also to keep track and preserve all biomaterials used to produce the image captures, since it is another way to ensure that these images can be reproduced.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethical aspects have already been addressed in the ethical section of the DoA as well as in the “D9.1 POPD Requirement number 3” document, since the intended use of human organs and tissues warrants serious consideration of ethical issues. As mentioned in that document, the partners will implement their research activities in full respect of the legal and ethical European / national / institutional requirements, and codes of practices. PIs from concerned laboratories (biobank and processing) have already full ethical clearance. Furthermore, the project management structure will include an Ethics Advisory Board (EAB), to provide guidance and ensure strict ethical governance relating to the documentation, application, material or any other aspect of the project that could have ethical implications.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question not answered.